expense or difficulties in cloning. Under these conditions it can be advantageous to sequence all or most clones in a library. In a subsequent round of evolution the deleterious mutations identified in the prior round can then be avoided altogether. In addition, all of the sequences present in the library can be sequenced if the number of clones to be assayed is small. It can be cost efficient to sequence even clones which have no activity because they help to improve the probability matrix. Sequencing using DNA or RNA arrays (Hyseq, Inc.) can be used.

After screening for a particular function, it can be determined which mutations affect that function. This information would help to understand the underlying mechanism of the functional protein. Furthermore, the next round of library construction can be focused on these positions and neighboring residues which produce the desired activity (*i.e.*, the constraint vector can be modified to better ensure functional proteins). The constraint vector can also be improved by determining the combinations of mutations that occur simultaneously in improved clones. These residues may interact and should be mutated simultaneously in subsequent rounds. Such synergistic mutations can be particularly important because they are almost impossible to identify by simple random mutagenesis.

Analysis of the library can also reveal the mutations that are missing from the unselected libraries. This could indicate toxicity, in addition to technical problems with library construction. If it is determined that an individual clone is toxic, such a polynucleotide or its encoded protein may find use as a drug or compound in which toxicity to bacteria is desired (assuming the library is constructed in *E. coli*). A related issue is the fitness distribution in the library. This can indicate the optimum mutation frequency for the library. The fitness distribution can also be used to compare various methods of calculating the probability matrix and the constraint vector, *i.e.*, the presence of continuous improvements of these methods.

Other useful products produced by the method of the invention include polynucleotides incorporating mutations identified through construction and screening of such libraries, vectors (including expression vectors) comprising such polynucleotides, host cells comprising such polynucleotides and/or vectors, and libraries of biological polymers, and libraries of host cells comprising and/or expressing such libraries of biological polymers.

VII. CORRELATION BETWEEN STRUCTURE AND FUNCTION OF PROTEIN MUTANTS

Statistical analyses of the correlation between structures and functions of molecules have been widely used to guide the optimization of small molecule drugs (quantitative structure activity relationship, or QSAR). One can differentiate between parameter-free approaches (for example Free, *J. Med. Chem.* (1964)) and methods which consider various physico-chemical parameters of the various substituents of a molecule (for example Carotti, *Chem Biol Interact* **67**:171 (1988)). See also, Goldman, et al., *Drug Development Research* **33**:125(1994) and Lahr, et al., *Proc. Nat'l Acad. Sci. USA* 96:14860 (1999). Either approach can be used for the libraries of the instant invention. In addition one can use algorithms based on the 3D structure of the protein of interest.

The amino acid sequence can be determined for variants that exhibit desired properties. The variants may each contain multiple mutations with respect to the parent molecule, and several variants may share one or more identical mutations while having other, nonshared mutations. The data mining task is to assign the degree to which individual mutations or combinations of mutations contribute to the observed improvement in properties, and to identify which pairs or groups of amino acids interact with each other (i.e. the observed measured property for the combined mutations is non-additive compared to the effect of the mutations individually). Methods for performing this data mining are known in the art; computer programs implementing suitable techniques are available (e.g., Spotfire).

VIII. CO-VARIATION AS A TOOL TO SELECT THE REGION TO BE MUTAGENIZED

Co-variation is the tendency of some residues to change simultaneously with other residues, *i.e.*, the residues are linked during evolution. These co-variant residues can be linked by structure and/or they may be linked by function. Once coupled residues have been identified, if one of the residues is found to be a candidate for mutation, the other residue can be assigned a higher probability of being a candidate as well. In this way, mutations which otherwise would not be obvious in a probability matrix or a constraint vector can be included. For further discussion of co-variation, see Gobel, et al., *Proteins* **18**:309 (1994); Jespers, et al., *J. Mol. Biol.* **290**:471 (1999); and Pazos, et al., *Comput. Appl. Biosci.* **13**:319 (1997).

VII. UTILITY OF THE LIBRARIES OF THIS INVENTION

While the utility of the libraries of this invention will be evident to one of skill in the art, the libraries will be particularly useful in preparation of enzymes or ligands with increased activity, enzymes or ligands with modified activity, proteins with increased stability, removal of immunogenic epitopes from useful proteins, improving expression levels of proteins, and improving grafting of domains or loops into proteins.

EXAMPLES

The following examples are set forth so as to provide those of ordinary skill in the art with a complete description of how to make and use the present invention, and are not intended to limit the scope of what is regarded as the invention. Efforts have been made to ensure accuracy with respect to numbers used (e.g., amounts, temperature, etc.) but some experimental error and deviation should be accounted for. Unless otherwise indicated, parts are parts by weight, temperature is degree centigrade and pressure is at or near atmospheric, and all materials are commercially available.

Example 1: Subtilisin With Novel Substrate Specificity

GG36 (savinase) is a subtilisin protease from *Bacillus lentus*. The goal of this Example is to generate mutants of the protease that possess a novel substrate specificity.

A published multiple sequence alignment of 124 subtilisin-like serine proteases (Siezen, et al., *Protein Science* **6**:501 (1997)) was recreated from a publicly available database (GENBANK), with the sequence labeled baalkp in the database being substituted with that of GG36. GG36 differs from baalkp by only one residue substitution. In baalkp, residue 87 is an asparagine while in GG36 a serine residue is found at the corresponding position. The GG36 amino acid sequence was used as the reference sequence, and those positions of the alignment for which the GG36 sequence had a gap character were deleted.

A profile for the alignment was generated using the method of Gribskov (Gribskov, *Proc. Nat'l Acad. Sci. USA* **84**:4355 (1987)) except that a mutation probability matrix was used in place of the log-odds matrix used by Gribskov. *See* Table 1. The mutation probability matrix gives the probabilities that a given amino acid will mutate to any another amino acid in a given evolutionary interval (Dayhoff, et al., *Atlas of Protein Sequence and Structure* (Natl. Biomed.